

Malware Classification

Coseinc Advanced research labs

Udi Shamir (tal0n)

udi@coseinc.com



Malware Classification

- During our work in the AML team we are analyzing hundreds of daily new and repeated malicious binaries, key loggers, phishing and other threats (Yeh, some are confidential :D) .
- while it is clear that without automation we could not even coverage 20% of our daily archive we had to write some new plugins for our CAMAL platform which perform the analysis section. the new plugins assist us detect new threats and better support our customers.

Malware Classification

existing products

- VxClass by Zynamics (Google)
-

Pros	Cons
use same algorithm as Bindiff	when new packer version arrive it fail
big signature database	pricey
intuitive	is now owned by Google :(

Malware Classification

existing products

- Yara, assist in identify static binary signatures, packers.
- Bindiff by Zynamics, compare and disassemble chunks of codes
- Binnavi by Zynamics, great for vulnerability detection and compare code flow

Malware Classification

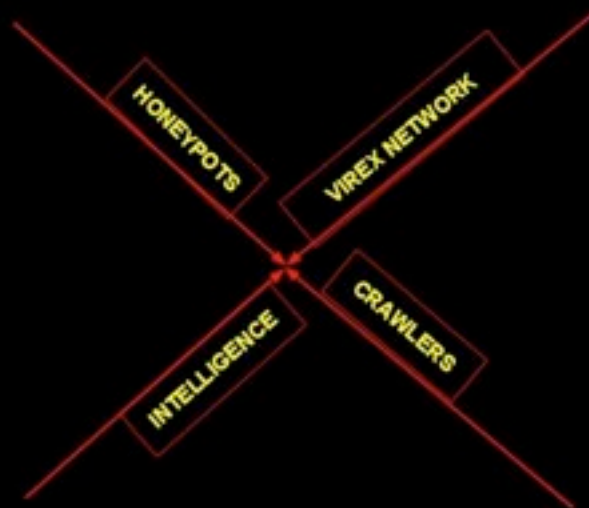
- categorized (labeling) malwares into families
- identifies behavioral patterns
- identify new variants

Categorization

- the categorization process helps us distinguish between:
- viruses, spywares, backdoors, trojans and worms.
- Detect new variants, targeted attacks

Categorization (Ingredients)

- Collecting Samples
- Auto analysis, sets
- Defining behavioral
- Clustering method



SAMPLE COLLECTION

- **VireX Network** is isolated network our partners are using to exchange new samples with the AML team. the exchange process provides usually good binary quality.
- **Honeypots** are deployed on tier1 providers separated by different geo location and are divided into 2 main categories: low/high interaction. the honeypots provides 5% of our daily malicious binaries collection.
- **Crawlers** we developed crawlers which scan related forums and leech binaries, the collection here is minimal but often provides new mutant.
- **Intelligence** ok , no need to add here :)

Auto Analysis (continue ...)

- Av vendors does not agree on standard naming convention
- Some vendors use better technique to differentiates between small variants in same family while others don't !
- we cannot rely on specific engine to help in the labeling process
- we decided to use one engine for the initial classification

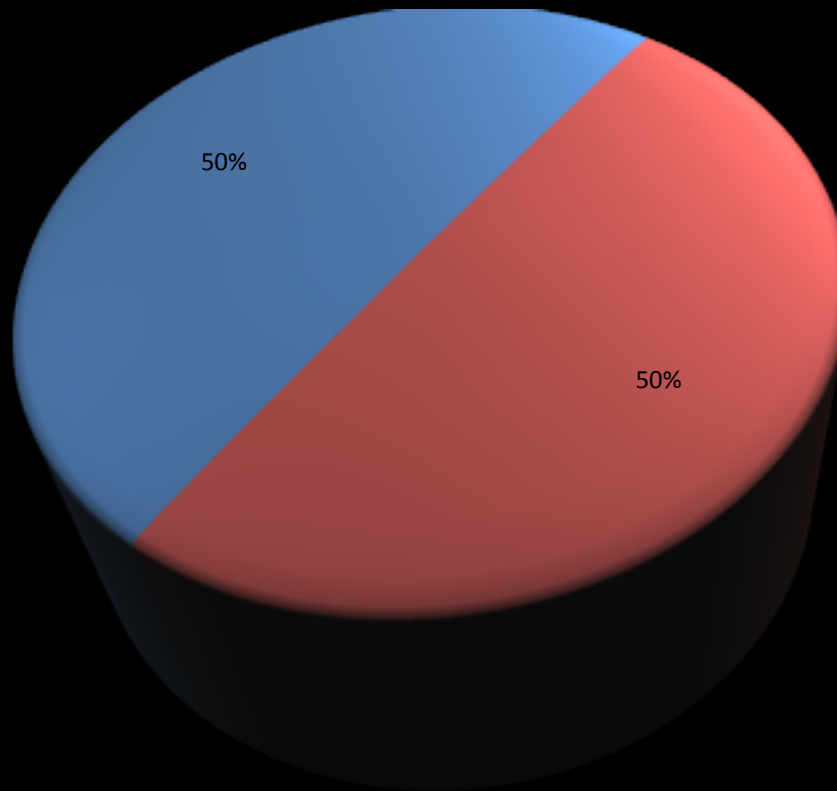
Auto Analysis (continue ...)

- We used legacy malware with 2 different hash from the same family.
- Malware signature exist in all engines

HASH	CLAMAV	FProt	Trend	McaFee
X	Blaster	msblast	msblast	Eploit-Dcom
Y	Dcom.Exploit	msblast	msblast	W32-Blaster-W

Auto Analysis (continue ...)

- The inspected AV engines provides around 50% consistency



Auto Analysis

- Dealing with the enormous samples we are collecting/ receiving each day is challenging and intractable.
- We are using the CAMAL framework to analyze each sample and to provide us with the initial datasets

Defining Behavioral

- Our reference data extracted from our large database located in the Camal AML
- The samples have been collected over the year taking average of 30GB a day.
- we uses one AV engine to perform the initial categorization for families
- Due to the skewed distribution we discarded classes with less then 10 samples

Defining Behavioral

Malware Classes	Quantity
Rbot	80
Sality	60
Virut	130
Storm	28

Defining Behavioral

- All malware are executed without user permission so we assume they are malicious.
- System call manipulations
- Loading Drivers
- Registry modification
- Modification of system files
- Geo location

Defining Behavioral

execute Camal on selected binaries

CAMAL PROFILER REPORT	
process creation	
"process", "created", "C:\WINDOWS\system32\rundll32.exe"	
write opcode	
"Write", "C:\Documents and Settings\admin\Start Menu\Programs\Internet Explorer.url"	
registry modification	
"registry", "SetValueKey" "{d669d1cf-0f47-11e0-928e-806d6172696f}\BaseClass"	
delete opcode	
"Delete", "C:\Documents and Settings\admin\Desktop\CaptureClient\binary.exe", "C:\RegTemp.txt"	
Geo Location	
Country Edition: BY, Belarus	

Defining Behavioral

execute Camal on selected binaries

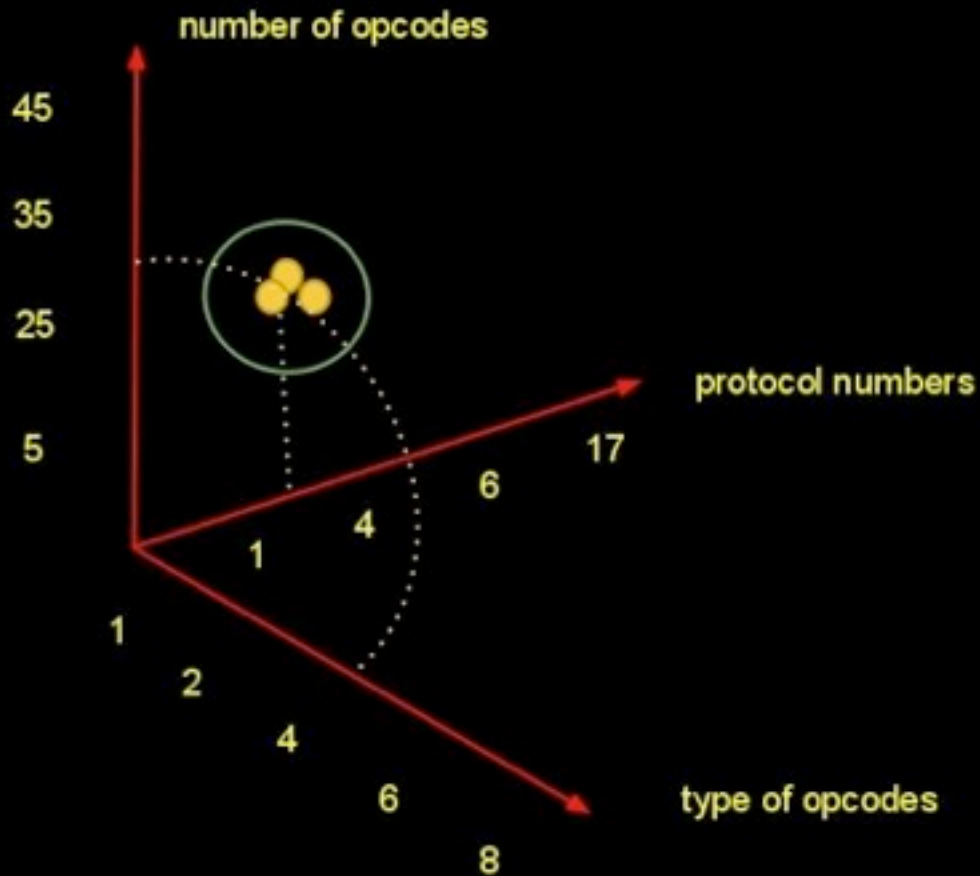
CAMAL PROFILER REPORT
number of opcodes (native system excluded)
28
type of opcodes (for specific list, see Camal DB reference)
4, 8, 10
drop zone
Country Edition: BY, Belarus
Protocols Numbers (for grouping list, see camal DB reference)
1, 4, 6
Geo Location

K-means

- One of the simplest learning algorithm
- Solve many known clustering problems

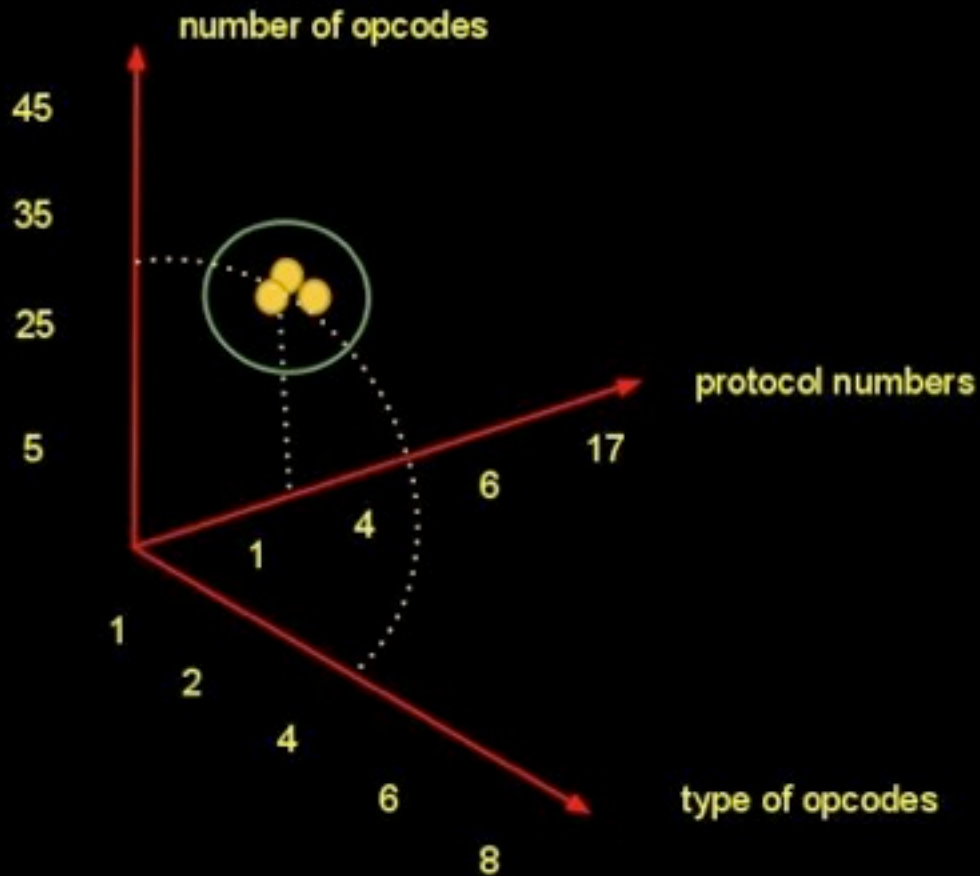
Categorization

K-means



Categorization

K-means



Summary

- Classification is important for better understanding how malwares operate
- Detect new outbreaks, provide early warning
- Assist in detecting new variants
- our method rely on emulation, hence we are not depend on unpackers (unless we are being detected)